

A. Experiment Details

In this section, we provide details about the implementation of our experiments. Please refer to our code at <https://github.com/Berkeley-NLP/Agent-Eval-Refine> for the official reference.

A.1. End-to-End Approach

We use `gpt-4-1106-vision-preview` through the OpenAI API and feed the image without resizing in “high-resolution”. We use a temperature of 0 and keep the other parameters at their default settings. The prompt templates for each environment are provided in Figures 12 and 13.

A.2. Modular Caption-then-Reason Approach

Collecting screenshots As described in Table 3, we constructed our dataset primarily through random subsampling from source datasets. However, for the iOS domain, due to limited online resources, we manually capture 50 extra screenshots in-house.

Action representation We represent actions as strings, e.g., `Type ``Hello```. This method leads to information loss when processing actions like clicks for pixel-based policies, as the click coordinates $[x, y]$ become meaningless when the image is represented by its textual description. We leave the task of more adequately transforming pixel-localized actions into textual forms for future work.

Collecting screenshot descriptions After obtaining the screenshots, we query GPT-4V (specifically, `gpt-4-1106-vision-preview` through the API) to get dense caption demonstrations. We manually fix or filter out ones with apparent errors. We use a temperature of 0 and keep the other parameters at their default settings. The prompt template is provided in Figure 15.

Finetuning Qwen-VL captioner The prompt template to query the finetuned Qwen-VL captioner is provided in Figure 16. We fine-tune the model over 3 epochs with a batch size of 72 and adamw optimizer [21], employing a cosine scheduler for learning rate adjustments starting from $1e-5$, a weight decay of 0.1, and a warmup ratio of 0.01. As shown in the prompt template, during both finetuning and at inference time, we provide text recognition results from the EasyOCR engine as an additional input to the model to reduce hallucination. We provide randomly-sampled model output examples in Figure 14.

Querying the reasoner After obtaining the descriptions of the screenshots and actions, we query the LM, either `Mixtral-8x7B-Instruct-v0.1` or `gpt-4-turbo-preview`. We use a temperature of 0

and keep the other parameters at their default settings. We provide the prompts to query the trajectory-level evaluator on Web and Android, and step-wise evaluator on iOS in Figures 17, 18, and 19 respectively.

A.3. Evaluation on WebArena

We directly use the `GPT4-0613 + CoT - v2` trajectories released by WebArena for evaluation. Confusion matrices for our evaluators’ predictions compared to the oracle evaluator are shown in Figure 4 (left).

A.4. Evaluation on Android

Emulator We use Android Studio’s built-in emulator to simulate a Pixel 4 with API version 33, and we develop a Python API for agent execution based on the `appium` package. We opt not to use `AndroidEnv` for Android emulation as it lacks support for typing actions.

Tasks The 120 evaluation tasks are evenly and randomly sampled from the General, WebShopping, and GoogleApps subsets of the Android-in-the-Wild test set (40 each) as shown in Listing 1. Note that we have excluded the Install and Single subsets. The install tasks require credit card information and are not safe to evaluate, while single-step tasks fall outside our focus on trajectory-level tasks.

Evaluation We use greedy decoding, i.e., temperature of 0, for all policies during evaluation. The confusion matrices comparing our evaluators with human judgments are presented in Figure 4 (right).

A.5. Refinement on WebArena

Reflexion We implement the Reflexion agent following the original paper [29]. The algorithm involves three key components: an Actor, an Evaluator, and a Self-Reflection module. The Actor generates thoughts and actions in text form based on the state observations it receives, where the actions are parsed into executable commands to step the environment for the next observation. The Evaluator assesses the quality of the outputs produced by the Actor. It computes a reward score based on how well the generated trajectory align with the expected outcomes of the task. If the evaluator assesses the task to be failed, the Self-Reflection model will be evoked to generate verbal reflections, which is stored in the agent’s memory and allows the agent to understand its past actions and their outcomes, facilitating improved decision-making in subsequent trials.

Implementation Details We use the DOM tree representation from the WebArena simulator as the environment observation. The LLM we use for Actor and the Self-Reflection is the `GPT-4-preview-1106` model, and the

prompts for these are shown in Listing 3 and 4 respectively. For the evaluator, we experiment with all the three variants proposed as well as the oracle evaluator from WebArena which is used for performance evaluation. Note that we use the webpage snapshot images instead of the DOM tree as the input to our evaluator.

A.6. Refinement on iOS

Emulator We use XCode’s built-in emulator to emulate an iPhone 13 device running iOS 16 and develop a Python API for the agent based on the [facebook/idb](#) package. We align its action space with that of the Android-in-the-Wild schema. Notably, since swiping up on the homescreen in Android means opening the AppDrawer, which can roughly be translated to swiping left or right on the homescreen, we bridge this domain gap by translating swiping up to swiping left or right 50% of the time during data collection and to the right 100% of the time during evaluation.

Tasks As shown in Listing 2, we design 132 task instructions that covers typical iOS device control tasks on Apple’s official apps, where 80 are used for data collection and training the agent, and 52 are hold out for testing. While we aimed to minimize the distribution difference between our tasks and that from the AitW dataset, the inherent differences between the platforms necessitated the inclusion of iOS-specific instructions, such as “Disable Siri’s access to Photos in Settings.”

Filtered BC Details We finetune the CogAgent model with its official code and apply LoRA [18], setting the parameters as follows: lora rank to 50, employing a cosine learning rate schedule, with a warmup proportion of 0.2, a learning rate (lr) of 0.00001, a batch size of 4 and train for 3000 steps. We use a relatively high temperature of 1.5 and topk=100 during data collection to improve diversity and greedy decoding (temperature=0) for all policies during evaluation.

Source	#	Domain
WebScreenshot [12]	128	Web
Mind2Web [11]	429	Web
AitW (train set) [27]	596	Android
GPT-4V in Wonderland [37]	60	iOS
In-house	50	iOS
Total	1,263	

Table 3. Sources of the screenshots dataset, including the number of screenshots paired with detailed descriptions and the original domain of the screenshots.

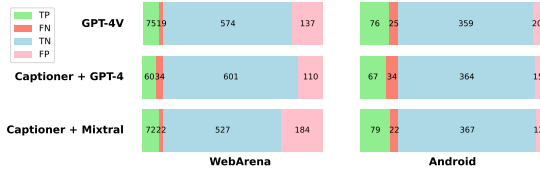


Figure 4. Confusion matrices for different evaluation systems against oracle evaluator or human judge on WebArena and Android.

Star the top one most starred repos in GitLab.

```
click [The A11Y Project/a11yproject.com]
go back
click [The A11Y Project/a11yproject.com]
stop
```

Response: The A11Y Project/a11yproject.com is already starred.

Thoughts: The user's intent was to star the top one most starred repo in GitLab, but the bot attempted to star a specific repository that was already starred. There is no indication that the repository in question is the top one most starred, and the bot did not perform any action to ensure it was interacting with the correct repository based on the user's intent. Therefore, the task is not completed successfully.

FAILURE

Figure 5. The evaluator detects a hard-to-find issue caused by a key missing step in the execution (i.e. ordering).

Post "lgtm" for the merge request related to semantic HTML post in the a11yproject project.

```
click [The A11Y Project/a11yproject.com']
click [Merge requests']
click [Add draft of What is Semantic HTML post]
scroll [down]
type [lgtm] at [Comment]
click at [Comment]
type [lgtm] at [Comment]
click at [Comment]
type [lgtm] at [Comment]
stop
```

Response: Early stop: Same typing action for 3 times.

Thoughts: The image shows that the comment "lgtm" has been posted twice by the user "Byte Blaze", which indicates that the bot successfully posted the comment as intended. However, the bot stopped early due to the same typing action being repeated three times, which is likely a safeguard against spamming. Since the user's intent was to post "lgtm" and it is visible in the merge request comments, the task is successful despite the early stop.

FAILURE

Figure 6. The evaluator correctly judge a task execution as successful despite the task ends with an exception.

Tell me who has made the most contributions (commits) to the Pytorch GAN project.

```
type [Pytorch GAN]
at [Search GitLab]
click [Erik Linder-Norén / PyTorch-GAN]
click [193 Commits]
stop
```

Response: Cannot determine.

Thoughts: The image shows the commits page of the PyTorch-GAN project on GitLab, but it does not provide a summary or a list that ranks contributors by the number of commits. The bot stopped with a "Cannot determine" response, which indicates it could not find the specific information requested by the user. The bot should have navigated to the 'Contributors' section to find the information about who has made the most contributions. Therefore, the task is not completed successfully.

FAILURE

Figure 7. The evaluator can also provide insight for possible corrective actions of failure executions.

Check if the walmart in pittsburgh can be reached in one hour by car from 5600 fifth avenue.

```
goto 'https://maps.google.com'
type '5600 Fifth Avenue, Pittsburgh, PA'
at [Search Google Maps]
click [Search nearby 5600 Fifth Ave]
type [Walmart] at [Search nearby 5600 Fifth Ave]
click [Search]
click [Get directions to Walmart Supercenter]
type [5600 Fifth Avenue, Pittsburgh, PA]
at [Choose starting point]
stop
```

Response: Yes, the Walmart in Pittsburgh can be reached in 16 minutes by car from 5600 Fifth Avenue, which is within the one-hour time frame.

Thoughts: The action history indicates that the agent performed a search for directions from "5600 Fifth Avenue, Pittsburgh, PA" to a Walmart in Pittsburgh. However, the final state of the webpage shows that the estimated travel time by car is actually "1 hour 43 minutes", which contradicts the bot's response to the user that stated a 16-minute travel time. This discrepancy suggests that the bot's response is incorrect, and the task of verifying if the Walmart can be reached within one hour by car has failed.

FAILURE

Figure 8. Captioning error where driving time is not captured, and the public transportation time is incorrectly captioned as driving time. The error is corrected if GPT4V is used.

Tell me the total cost of my latest cancelled order?

```
click [My Account]
click [My Orders]
stop
```

Response: \$231.54.

Thoughts: The user's intent is to know the total cost of their latest cancelled order. From the action history, we see that the bot successfully navigated to "My Account" and "My Orders" pages. The final state of the webpage shows a list of orders with details. The latest cancelled order has a price of \$231.54, which matches the bot's response. Therefore, it seems that the bot has successfully completed the task.

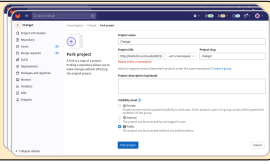
SUCCESS

Figure 9. Mixtral failed to identify that the price is not from a cancelled order from a cluttered webpage. The error is corrected if GPT4/GPT4V is used.

Help me Fork ChatGPT.

type [ChatGPT]
 at [Search GitLab]
 click [Convex Egggart / Chatgpt]
 click [Fork]
 click [Fork project]
 stop

Response: Forking ChatGPT is in progress.



Thoughts: The action history indicates that the agent has performed the necessary steps to initiate the forking of the ChatGPT repository on GitLab. The final state of the webpage shows that the agent is on a "Fork project" form, which is consistent with the forking process. The bot's response to the user confirms that the forking process has been initiated. There is no indication of an error or incomplete action, so we can assume the task was successful unless there is evidence to the contrary.

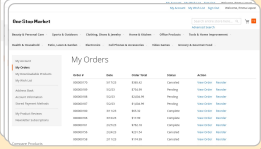
SUCCESS

Figure 10. All the evaluation models fail in identifying the failure because they miss to see the error message which indicates the failure of forking.

How much I spent on food-related shopping during March 2023?

1: click [My Account]
 2: click [My Orders]
 3: click [View Order]
 4: click [My Orders]
 5: click [View Order]
 6: click [My Orders]
 stop

Response: \$83.31.



Thoughts: The bot's response only includes a single order total, but the user's intent was to know the total amount spent on food-related shopping for the entire month of March 2023. The image shows multiple orders in March, and the bot should have summed up all completed and not canceled orders to provide an accurate total. Therefore, the task has failed.

FAILURE

Figure 11. Though our approach successfully identifies that the task has failed, the reasoning is incorrect. The error is caused by not filtering the orders for food-related only, instead of not summing up the total price.

System Prompt

You are an expert in evaluating the performance of a web navigation agent. The agent is designed to help a human user navigate a website to complete a task. Given the user's intent, the agent's action history, the final state of the webpage, and the agent's response to the user, your goal is to decide whether the agent's execution is successful or not. There are three types of tasks: 1. Information seeking: The user wants to obtain certain information from the webpage, such as the information of a product, reviews, map info, comparison of map routes, etc. The bot's response must contain the information the user wants, or explicitly state that the information is not available. Otherwise, e.g. the bot encounters an exception and respond with the error content, the task is considered a failure. Besides, be careful about the sufficiency of the agent's actions. For example, when asked to list the top-searched items in a shop, the agent should order the items by the number of searches, and then return the top items. If the ordering action is missing, the task is likely to fail. 2. Site navigation: The user wants to navigate to a specific page. Carefully examine the bot's action history and the final state of the webpage to determine whether the bot successfully completes the task. No need to consider the bot's response. 3. Content modification: The user wants to modify the content of a webpage or configuration. Carefully examine the bot's action history and the final state of the webpage to determine whether the bot successfully completes the task. No need to consider the bot's response.

IMPORTANT Format your response into two lines as shown below:
 Thoughts: <your thoughts and reasoning process>
 Status: ``success" or ``failure"

User Prompt

User Intent: intent
 Action History: {last actions}
 The last snapshot of the web page is shown in the image.
 Bot response to the user: {response if response else ``N/A"}.

Figure 12. Prompts to query the VLM Evaluator in WebArena.

System Prompt

You are an expert in evaluating the performance of an android navigation agent. The agent is designed to help a human user navigate the device to complete a task. Given the user's intent, and the final state of the screen, your goal is to decide whether the agent has successfully completed the task or not.

IMPORTANT Format your response into two lines as shown below:
 Thoughts: <your thoughts and reasoning process>
 Status: "success" or "failure"

User Prompt

User Intent: {intent}

Action History:
 {last.actions}

The last snapshot of the screen is shown in the image.

Bot response to the user: {response if response else "N/A"}.

Figure 13. Prompts to query the VLM Evaluator in Android

GUI Interface Description (Continued)

- ## Top Navigation Bar
 - **DASHBOARD** (selected)
 - SALES
 - CATALOG
 - CUSTOMERS
 - CONTENT
 - MARKETING
 - STORES
 - SYSTEM
 - Find Partners | Extensions
- ## Dashboard Section
 - ### Left Pane - Filter Options
 - Scope dropdown menu with "All Store Views" selected
 - Reload Data button
 - ### Main Content Area
 - #### Advanced Reporting Card
 - Title: "Advanced Reporting"
 - Subtitle: "Gain new insights and take command of your business's performance, using our dynamic product, order, and customer reports tailored to your customer data."
 - Checkbox option: "Chart is disabled. To enable the chart, click here."
 - #### Lifetime Sales Table
 - Header row with columns labeled: Customer Name, Items Total, Product, Price, Quantity
 - Items Total
 - Product
 - Price
 - Quantity
 - #### Bottom Navigation Footer
 - icon for customers with badge count overlay indicating "3"
 - (Note: The description provided captures visible elements on the interface without identifying any real-life individuals or making assumptions about their roles or actions.)

(a) WebArena Example

GUI Interface Description

- ## Top Navigation Bar
 - Hugging Face logo on the left
 - Search bar labeled "Search models, datasets, ..." with a magnifying glass icon
 - Menu items: Models, Datasets, Spaces, Posts, Docs, Pricing
 - User icon with dropdown arrow indicating more options
- ## Project Overview Section
 - ### Agent-Eval-Refine Header
 - Labelled as "Upgrade to Enterprise"
 - ### New Button
 - Located next to Activity Feed button
 - ### Organization Card
 - #### None Defined Yet Section
 - Text stating "None defined yet."
 - #### Organization Title
 - "Autonomous Evaluation and Refinement of Digital Agents" with an information (i) icon
 - #### Team Members List
 - A list of individuals or entities involved with this project, including names like "Jiayi Pan", "Yichi Zhang", etc.
 - #### Project Information Link
 - Labeled "Project Page: TODO"
 - Hyperlink pointing to page titled "TODO"
 - #### Paper Information Link
 - Labeled "Paper: TODO"
 - Hyperlink pointing to paper titled "TODO"

(b) Random Real Website

GUI Interface Description

- ## Top Status Bar
 - Time display: "1:42"
 - Cellular signal strength indicator
 - Wi-Fi signal icon
 - Battery status and level indicator
- ## Main Content Area
 - ### Contacts Header
 - Title: "Contacts"
 - ### Contact Information Section
 - #### Contact Card Example
 - Avatar placeholder with circular symbol and initial "KB" inside it
 - Name: "Kate Bell"
 - Job Title: "Producer | Creative Consulting"
 - #### Action Buttons
 - Message button: Blue square with white message bubble icon
 - Call button: Red square with telephone receiver icon
 - Mail button: Grey square with envelope icon
 - #### Contact Details Row
 - Mobile number field: "(555) 564-8583"
 - Work phone number field: "(415) 555-3695"
 - #### Additional Contact Info Column
 - Email address field: "kate-bell@mac.com"
 - Homepage URL field: "http://www.icloud.com"
 - Work location and address: "165 Davis Street • Hillsborough CA 94010"
 - #### Birthday Field
 - Label: "birthday"
 - Date of Birth: "January 20, 1978"
 - ### Bottom Navigation Bar
 - Back navigation arrow (left)
 - Edit text input field labeled "Edit"
 - Add new contact (+) icon on the right side

(c) iOS Example

GUI Interface Description

- ## Status Bar
 - Time display (12:54)
 - Network signal indicator
 - Wi-Fi icon
- ## App Interface Header
 - Search bar with magnifying glass icon and placeholder text "Search apps & games"
 - Profile or user account icon on top right corner labeled as "X"
- ## Main Content Area
 - Section title "For you" with dropdown arrow indicating more options
 - Sub-section "Top charts" with circular refresh button
 - Sub-section "Kids" with child's hand icon
 - Sub-section "Categories" with bookmark icon
- ## Featured Promotion Banner
 - Image depicting an abstract background with floating shapes and patterns, including what appears to be a star shape at center
 - Promotional text overlay saying "Special event"
 - Headline "Embrace new possibilities in 2024"
 - Additional subtitle "New year, new you"
- ## Sponsored - Suggested for you
 - Label "Sponsored - Suggested for you"
 - List of Recommended Apps
 - TikTok app card
 - Logo (orange sticking out emoji face)
 - Category label "Social • Networking"
 - Rating stars showing four-and-a-half out of five
 - Version number "8.3.4"
 - TextNow app card
 - Company logo (stylized T symbol)
 - Category label "Communication • Call management"
 - Rating stars showing four-and-a-half out of five
 - Version number "4.5.6"
 - Temu app card
 - Company logo (stylized T symbol)
 - Category label "Shopping • Online marketplace"
 - Rating stars showing four-and-a-half out of five
 - Version number "4.7.8"
- ## Bottom Navigation Menu
 - Games tab highlighted (indicated by color)
 - Apps tab
 - Offers tab
 - Books tab

The interface is that of a mobile application store page, likely Google Play Store given the design elements visible in the image provided. The content displayed suggests popular applications available for download within this platform.

(d) Android Example

Figure 14. Example outputs from the captioner model across different environments. Examples are selected at random without cherry-picking.

User Prompt
 You are an advanced GUI captioner. Please describe this GUI interface in details and don't miss anything. Your response should be hierarchical and in Markdown format. Don't do paraphrase. Don't wrap your response in a code block.

Figure 15. Prompts to query GPT-4V for collecting dense captions

User Prompt
 Please describe the screenshot above in details.
 OCR Result:
 {ocr_result}

Figure 16. Prompts to query the fine-tuned captioner for dense captioning.

System Prompt
 You are an expert in evaluating the performance of a web navigation agent. The agent is designed to help a human user navigate a website to complete a task. Given the user's intent, the agent's action history, the final state of the webpage, and the agent's response to the user, your goal is to decide whether the agent's execution is successful or not.

There are three types of tasks:

1. Information seeking: The user wants to obtain certain information from the webpage, such as the information of a product, reviews, map info, comparison of map routes, etc. The bot's response must contain the information the user wants, or explicitly state that the information is not available. Otherwise, e.g. the bot encounters an exception and respond with the error content, the task is considered a failure. Besides, be careful about the sufficiency of the agent's actions. For example, when asked to list the top-searched items in a shop, the agent should order the items by the number of searches, and then return the top items. If the ordering action is missing, the task is likely to fail.
2. Site navigation: The user wants to navigate to a specific page. Carefully examine the bot's action history and the final state of the webpage to determine whether the bot successfully completes the task. No need to consider the bot's response.
3. Content modification: The user wants to modify the content of a webpage or configuration. Carefully examine the bot's action history and the final state of the webpage to determine whether the bot successfully completes the task. No need to consider the bot's response.

IMPORTANT
 Format your response into two lines as shown below:
 Thoughts: <your thoughts and reasoning process>
 Status: ``success" or ``failure"

User Prompt
 User Intent: {intent}
 Action History:
 {last.actions}
 The detailed final state of the webpage:
 ```md  
 {cap}  
 ```

Figure 17. Prompts to query the LM reasoner in WebArena

System Prompt
 You are an expert in evaluating the performance of an android navigation agent. The agent is designed to help a human user navigate the device to complete a task. Given the user's intent, and the state of the screen, your goal is to decide whether the agent has successfully completed the task or not.
 IMPORTANT Format your response into two lines as shown below:
 Thoughts: <your thoughts and reasoning process>
 Status: "success" or "failure"

User Prompt
 User Intent: {intent}

Action History:
 {last.actions}

The detailed final state of the screen:
 ```md  
 {cap}  
 ```

Figure 18. Prompts to query the LM Reasoner for trajectory-level evaluation in Android

System Prompt
 You are a GUI Trajectory Evaluator. Your task is to observe a bot's action within a graphical user interface (GUI) and classify its behavior into one of four categories based on its progress towards a specified goal. The categories are:

1. "towards-the-goal" - The bot is moving closer to achieving the goal.
2. "not-sure" - It's unclear if the bot's actions are helping reach the goal.
3. "goal-reached" - The bot has successfully completed the goal.
4. "away-from-the-goal" - The bot's actions are diverting it from the goal.

Please format your response as follows:

Thoughts: [Explain your reasoning here]
 Response: "towards-the-goal", "not-sure", "goal-reached", or "away-from-the-goal"

Here are some example responses:

Example 1:
 Thoughts: The goal is to 'set an alarm at 8:00 am.' Initially, the bot is on the home screen. After a tap action, it navigates to the alarm app, indicating progress towards the goal.
 Response: "towards-the-goal"

Example 2:
 Thoughts: The goal is to 'buy the latest iPhone on Amazon.' The bot starts at the checkout page on Amazon. After a tap action, the screen shows a successful purchase, signifying that the goal has been reached.
 Response: "goal-reached"

Example 3:
 Thoughts: The goal is to 'show me the weather in New York.' The bot begins on London's weather page. After pressing 'home', it returns to the home screen, moving away from the goal.
 Response: "away-from-the-goal"

Example 4:
 Thoughts: The goal is to 'buy some coffee on the Starbucks app.' The bot begins on the Amazon app. After pressing 'back,' it moves to the home screen, which is a prerequisite for opening the Starbucks app.
 Response: "towards-the-goal"

Example 5:
 Thoughts: The goal is to 'open YouTube.' The bot begins on the home screen. After a swipe, it appears to remain on the same page, suggesting no progress towards the goal.
 Response: "not-sure"

Note:
 You should be extra-careful when assigning "goal-reached" or "towards-the-goal" labels. If you are unsure, please select "not-sure" instead.

User Prompt
 Goal: {intent}
 Original State:
 ```md  
 {current.state} ```

State after action: "{action}":  
 ```md  
 {next.state}
 ```

Figure 19. Prompts to query the LM Reasoner in iOS for per-step evaluation.

**Listing 1. Evaluation tasks sampled from Android-in-the-Wild test set**

```

--- General ---
Open the calendar
Play the new Katy Perry video on YouTube
What's on the menu at Papa Murphy's?
How much does a 2 bedroom apartment rent for in
 ↳ Philadelphia?
What's the price of the Galaxy phone on eBay?
Open a new private window in Chrome
How do I get to the nearest Target?
Install the Twitter app
Search for flights from NYC to Mexico city
What is the capital of China?
What's US dollar exchange rate against the South Korean
 ↳ Won?
What is the speed of a train?
What is the capital of Norway?
What's the news in Peru?
How do I get to the nearest IKEA?
What's a good restaurant in Miami?
How much does a 3 bedroom apartment rent for in Houston?
What's the weather like in Chicago?
Set an alarm for 6pm
What is the capital of Mexico?
What's the news in Cambodia?
What's a good restaurant in Chicago?
What is the capital of Japan?
Search for flights from Barcelona to Mexico city
What's on the menu at Olive Garden?
What's the news in India?
What is the capital of Canada?
Search for a new nail polish
Open a new Chrome tab
What's the latest news in planetary science?
How much does a 2 bedroom apartment rent for in Boston?
Search for hotels in Las Vegas
Where can I buy a nice beach towel?
What's the price of the LG TV?
Search for a new highlighter
What's the latest video from GameXplain?
Search for top rated sushi restaurants on Maps
What's the weather like in Johannesburg?
Search for hotels in London
Install the Google app
--- Web Shopping ---
Search for a new 65" TV at Best Buy
Add "sony triple a" to the cart on amazon
Search for "macbook pro 15 inch" on walmart.com, select
 ↳ the first entry, and add it to the cart.
Look up the best gaming headphones on Best Buy
Add bose quietcomfort 35 to the cart on costco.com
Add macbook pro to the cart on amazon.com
Add "dell xps" to the cart on bestbuy, then select
 ↳ checkout.
Clear the shopping cart on newegg.com. Search for "razer
 ↳ naga" on newegg.com, select the first entry,
 ↳ and add it to the cart.
View the shopping cart on ebay. Search for "usb-a" on
 ↳ ebay, select the first entry, add it to the cart
 ↳ , then select checkout.
Search for "rayovac triple a" on walmart, select the
 ↳ first entry, and add it to the cart.
Clear the cart on newegg. Search for duracell triple a
 ↳ on newegg, select the first entry, add it to the
 ↳ cart, then select checkout.
Add bose quietcomfort 35 to the cart on amazon.com
Search for jbl flip 4 on ebay.com, select the first
 ↳ entry, add it to the cart, then select checkout.
Clear the cart on newegg.com. Add usb-a to the cart on
 ↳ newegg.com, then select checkout.
Clear all items from cart on costco. Search for bose
 ↳ quietcomfort 35 on costco, select the first
 ↳ entry, and add it to the cart.
Search for the best books of all time on Goodreads
Clear the cart on target.com. Search for "logitech g502"
 ↳ on target.com, select the first entry, and add
 ↳ it to the cart.
Clear the cart on costco.com. Search for usb-c to usb-a

```

```

 ↳ on costco.com, select the first entry, and add
 ↳ it to the cart.
Search for duracell triple a on newegg.com, select the
 ↳ first entry, and add it to the cart.
Show the shopping cart on amazon. Add dell alienware to
 ↳ the cart on amazon, then select checkout.
Search for acer predator on ebay.com, select the first
 ↳ entry, and add it to the cart.
Show the shopping cart on ebay.com. Search for "razer
 ↳ blackwidow" on ebay.com, select the first entry,
 ↳ add it to the cart, then select checkout.
Add razer nari to the cart on target, then select
 ↳ checkout.
Add bose soundlink to the cart on amazon.com
Empty the shopping cart on newegg.com. Add jbl charge 4
 ↳ to the cart on newegg.com, then select checkout.
Clear the shopping cart on costco.
Clear all items from cart on bestbuy.com. Search for usb
 ↳ -a on bestbuy.com, select the first entry, and
 ↳ add it to the cart.
Clear the shopping cart on amazon.com. Search for razer
 ↳ blade on amazon.com, select the first entry, add
 ↳ it to the cart, then select checkout.
Add macbook air to the cart on target
Search for "corsair k70" on target.com, select the first
 ↳ entry, add it to the cart, then select checkout
 ↳ .
View the shopping cart on bestbuy. Search for "logitech
 ↳ g pro" on bestbuy, select the first entry, add
 ↳ it to the cart, then select checkout.
View the shopping cart on bestbuy.com. Search for
 ↳ panasonic triple a on bestbuy.com, select the
 ↳ first entry, add it to the cart, then select
 ↳ checkout.
Clear the shopping cart on target.com. Add macbook air
 ↳ to the cart on target.com, then select checkout.
Search for razer blade on bestbuy.com, select the first
 ↳ entry, and add it to the cart.
Clear the cart on target. Add "usb-c to usb-a" to the
 ↳ cart on target
Search for the best books on Goodreads
Search for a 3d printer on aliexpress
Search for duracell triple a on costco.com, select the
 ↳ first entry, add it to the cart, then select
 ↳ checkout.
Add lg ultragear to the cart on bestbuy.com
Clear all items from cart on walmart.com. Add macbook
 ↳ pro to the cart on walmart.com, then select
 ↳ checkout.
--- Google Apps ---
Do I have any events tomorrow?
toggle data saver in the chrome app
turn off airplane mode
Open calendar and show me the first week of next month
Go to location settings
create a new album in the google photos
Open calendar and show me the fourth week of next month
What's on my calendar today?
check android version
add a label to a message in the gmail app
Open the calendar and show me this week's events?
turn on the 24-hour format for clock
empty trash in google photos
turn on showing notifications on the lock screen
turn on bluetooth scan
toggle notification dots
set the timer
Open Wikipedia
open chrome and create a bookmark for the current page
change the upload size in google photos
Show me popular videos on Youtube
Go to settings
change the clock display to analog
Search for Italian restaurants on Maps
turn pop-ups on in chrome
turn vacation reply on in the gmail app
turn on wifi
Search for pizza restaurants on Maps
Check the weather

```



```

toggle sleep mode
turn notification dots off
turn off translation in the chrome app
all mails in gmail
Go to privacy settings
Is it going to rain today?
toggle wifi
Is it going to rain this weekend?
Open calendar and show me the second week of next month
see creations saved in the google photos
Open internet settings

```

## Listing 2. Data collection and evaluation tasks for iOS experiments

```

--- Train Tasks ---
Open the Reminder App
Open the Calendar App
Open the Map App
Open the Contacts App
Open Safari
Open the Wallet App
Open Messages
Open Health App
Open Files App
Open Shortcuts App
Open Freeform App
Open Watch App
Open General Page in Settings App
Find the About Page in Settings App
Find the iOS version of this device in Settings
Find the Serial Number of this device in Settings
Disable Auto-Correction for Keyboard in Settings
Disable Smart Punctuation for Keyboard in Settings
Disable Slide to Type for Keyboard in Settings
Disable Caps Lock for Keyboard in Settings
Change the Temperature Degree to Celsius in Settings/
 ↳ General
Change Safari Search Engine to Yahoo in Settings
Change Safari Search Engine to Bing in Settings
Change Safari Search Engine to Baidu in Settings
Change Safari Search Engine to Sogou in Settings
Change the preferred language of Safari to Chinese in
 ↳ Settings
Change the preferred language of Safari to Japanese in
 ↳ Settings
Change the preferred language of Safari to Russian in
 ↳ Settings
Change the preferred language of Safari to Thai in
 ↳ Settings
Change the preferred type of travel of Maps to Driving
 ↳ in Settings
Change the preferred type of travel of Maps to Transit
 ↳ in Settings
Change the preferred type of travel of Maps to Cycling
 ↳ in Settings
Disable location access of Maps in Settings
Disable Siri's access to Maps in Settings
Disable Siri's access to Health in Settings
Disable Siri's access to News in Settings
Create a reminder to "Buy a Birthday Gift for Alice"
Create a reminder of "Dinner with Peter"
Create a reminder of "Lunch with Dave"
Create a reminder of "Christmax Eve"
Send a message to Kate Bell say "How are you doing?"
Send a text message using the Messages app to the first
 ↳ contact saying "Happy New Year!"
Use the Reminders app to set a reminder for "Water the
 ↳ plants" for tomorrow.
Use the Reminders app to set a reminder for "Wash
 ↳ Clothes" for tomorrow.
Show me the latest image I took in the Photos
Show me the photos in the album "Recent"
Show me the photos in the album "Favorite"
Use Maps to find the nearest McDonald's.
Use Maps to find the nearest grocery store.
Use Maps to find the nearest gas station.
Use Maps to find the nearest metro station.

```

```

Create an event called "dinner" on calendar for today
Create an event called "dinner" on calendar for tomorrow
Create an event called "lunch" on calendar for today
Create an event called "Meeting with Joey" on calendar
 ↳ for today 8pm
Create an event called "Meeting with Anne" on calendar
 ↳ for tomorrow 2pm
Create an event called "Meeting with Simon" on calendar
 ↳ for today 8am
Use the Contacts app to find David Taylor's phone number
 ↳ .
Use the Contacts app to find Kate Bell's phone number.
Use the Contacts app to find Daniel's phone number.
Use the Contacts app to find Anna Haro's birthday.
Use the Contacts app to find David Taylor's birthday.
Use the Contacts app to find David Taylor's address.
Use the Contacts app to find Kate Bell's address.
Use the Contacts app to find Daniel's address.
Use the Contacts app to find John Appleseed's address.
Create a new Contact called Bill.
Create a new Contact called Dan with phone number
 ↳ 8888-8888.
Create a new Contact called Kelly with phone number
 ↳ 1234-5678.
Use the Calendar app to add a new all-day event called "
 ↳ On Vacation" at tomorrow.
Use the Calendar app to add a new all-day event called "
 ↳ On Vacation" for yesterday.
Open Files and create a new folder named "Work Documents
 ↳ " at "On My iPhone".
Open the News app and open the top-most news article.
Open Safari and show me MIT's wikipedia page.
Open Safari and show me Stanford's wikipedia page.
Open Safari and show me CMU's wikipedia page.
Open Safari and show me Apple's wikipedia page.
Open Safari and show me Microsoft's wikipedia page.
Open Safari and show me Amazon's wikipedia page.

--- Eval Tasks ---
Open Settings
Open News
Open the app Watch
Open the Settings and open settings for Safari
Find the nearest gas station on Map
Create a reminder of "make dinner"
Create a reminder of "grocery"
Create a reminder of "finish homework"
Send a message to Kate with "Hi"
Send a message to John Appleseed with "How are you doing
 ↳ "
Find the nearest gas station with Maps
Use Maps to find the nearest Target
Use Maps to find the nearest CVS
Use Contacts to find John Appleseed's birthday
Use Contacts to find Kate Bell's phone number
Find Anna Haro's phone number with the Contacts app
Find the app Freeform and open it
Find the app Shortcuts and open it
Open Settings and open About page
Find John Appleseed's phone number with contacts
Find Kate Bell's phone number with contacts
Show me an image in Photos
Show me all images in Photos
Open maps for me
add a reminder to make dinner
add a reminder to do grocery
find the nearest Nike store on maps
send a message to Kate Bell with the message app
send a message to John with message app
find the nearest metro station on maps
Use Contacts to check out John Appleseed's home address
Open the Settings and open Accessibility
Find the Model Number of this device
Disable Character Preview for Keyboard in Settings
Change Safari Search Engine to DuckDuckGo in Settings
Change the preferred language of Safari to Turkish in
 ↳ Settings
Change the preferred type of travel of Maps to Walking
 ↳ in Settings

```

Disable Siri's access to Photos in Settings  
 Create a reminder of "Christmax gift for Kitty"  
 Use the Reminders app to set a reminder for "Finish  
 → homework" for tomorrow.  
 Send a message to John say "How are you doing?"  
 Show me all the images I have in Photos  
 Show me the photos in the album "Favorite"  
 Use Maps to find the nearest Burger King.  
 Create an event called "dinner" on calendar for tomorrow  
 Create an event called "Meeting" on calendar for today 1  
 → pm  
 Use the Contacts app to find John Appleseed's phone  
 → number.  
 Use the Contacts app to find Kate Bell's birthday.  
 Use the Contacts app to find Anna Haro's address.  
 Create a new Contact called Simon with phone number  
 → 1234-5678.  
 Use the Calendar app to add a new all-day event called "  
 → On Vacation" for today.  
 Open Safari and show me UC Berkeley's wikipedia page.

### Listing 3. Prompts for Reflexion Agent to Take Actions

```
prompt = {
 "intro": """"You are an autonomous intelligent agent
 → tasked with navigating a web browser. You will
 → be given web-based tasks. These tasks will be
 → accomplished through the use of specific actions
 → you can issue.

 Here's the information you'll have:
 The user's objective: This is the task you're trying to
 → complete.
 The current web page's accessibility tree: This is a
 → simplified representation of the webpage,
 → providing key information.
 The current web page's URL: This is the page you're
 → currently navigating.
 The open tabs: These are the tabs you have open.
 The previous action: This is the action you just
 → performed. It may be helpful to track your
 → progress.

 The actions you can perform fall into several categories
 → :

 Page Operation Actions:
 'click [id]': This action clicks on an element with a
 → specific id on the webpage.
 'type [id] [content] [press_enter_after=0|1]': Use this
 → to type the content into the field with id. By
 → default, the "Enter" key is pressed after typing
 → unless press_enter_after is set to 0.
 'hover [id]': Hover over an element with id.
 'press [key_comb]': Simulates the pressing of a key
 → combination on the keyboard (e.g., Ctrl+v).
 'scroll [direction=down/up]': Scroll the page up or down
 → .

 Tab Management Actions:
 'new_tab': Open a new, empty browser tab.
 'tab_focus [tab_index]': Switch the browser's focus to a
 → specific tab using its index.
 'close_tab': Close the currently active tab.

 URL Navigation Actions:
 'goto [url]': Navigate to a specific URL.
 'go_back': Navigate to the previously viewed page.
 'go_forward': Navigate to the next page (if a previous '

 → go_back' action was performed).

 Completion Action:
 'stop [answer]': Issue this action when you believe the
 → task is complete. If the objective is to find a
 → text-based answer, provide the answer in the
 → bracket. If you believe the task is impossible
 → to complete, provide the answer as "N/A" in the
 → bracket.
```

Homepage:  
 If you want to visit other websites, check out the  
 → homepage at <http://homepage.com>. It has a list  
 → of websites you can visit.  
<http://homepage.com/password.html> lists all the account  
 → name and password for the websites. You can use  
 → them to log in to the websites.

To be successful, it is very important to follow the  
 → following rules:

1. You should only issue an action that is valid given  
 → the current observation
2. You should only issue one action at a time.
3. You should follow the examples to reason step by step  
 → and then issue the next action.
4. Generate the action in the correct format. Start with  
 → a "In summary, the next action I will perform  
 → is" phrase, followed by action inside '''''.  
 → For example, "In summary, the next action I will  
 → perform is ''click [1234]'''. Including your  
 → thinking process before giving the action is  
 → encouraged.
5. Issue stop action when you think you have achieved  
 → the objective. Don't generate anything after  
 → stop.'''',

"examples": [  
 (  
 ""OBSERVATION:  
 [1744] link 'HP CB782A#ABA 640 Inkjet Fax Machine (  
 → Renewed)'  
 [1749] StaticText '\$279.49'  
 [1757] button 'Add to Cart'  
 [1760] button 'Add to Wish List'  
 [1761] button 'Add to Compare'  
 URL: [http://onestopmarket.com/office-products/office-  
 → electronics.html](http://onestopmarket.com/office-products/office-electronics.html)  
 OBJECTIVE: What is the price of HP Inkjet Fax Machine  
 MEMORY: None  
 PREVIOUS ACTION: None''''',  
 "Let's think step-by-step. This page list the  
 → information of HP Inkjet Fax Machine, which is  
 → the product identified in the objective. Its  
 → price is \$279.49. I think I have achieved the  
 → objective. I will issue the stop action with the  
 → answer. In summary, the next action I will  
 → perform is ''stop [\$279.49]''',  
 ),  
 (  
 ""OBSERVATION:  
 [164] textbox 'Search' focused: True required: False  
 [171] button 'Go'  
 [174] link 'Find directions between two points'  
 [212] heading 'Search Results'  
 [216] button 'Close'  
 URL: <http://openstreetmap.org>  
 OBJECTIVE: Show me the restaurants near CMU  
 MEMORY: I first searched "CMU" and then searched for "  
 → restaurants". It looks like the search results  
 → are different from what I expected, as the  
 → search results are about the restaurants in  
 → another city. I think I should go back to the  
 → previous page and search for "restaurants near  
 → CMU".  
 PREVIOUS ACTION: None''''',  
 "Let's think step-by-step. This page has a search  
 → box whose ID is [164]. According to the  
 → nominatim rule of openstreetmap, I can search  
 → for the restaurants near a location by \  
 → restaurants near\". I can submit my typing by  
 → pressing the Enter afterwards. In summary, the  
 → next action I will perform is ''type [164] [  
 → restaurants near CMU] [1]''',  
 ),  
 ],  
 "template": ""OBSERVATION:  
 {observation}  
 URL: {url}  
 OBJECTIVE: {objective}  
 MEMORY: {memory}

```
PREVIOUS ACTION: {previous_action}""",
 "meta_data": {
 "observation": "accessibility_tree",
 "action_type": "id_accessibility_tree",
 "keywords": ["url", "objective", "observation", "
 ↳ previous_action", "memory"],
 "prompt_constructor": "ReflexionPromptConstructor",
 "answer_phrase": "In summary, the next action I will
 ↳ perform is",
 "action_splitter": "```"
 },
}
```

#### Listing 4. Prompts for Reflexion Agent to Reflect

```
prompt = {
 "intro": ""You are an autonomous intelligent agent
 ↳ tasked with navigating a web browser. You will
 ↳ be given web-based tasks. These tasks will be
 ↳ accomplished through the use of specific actions
 ↳ you can issue.

 Here's the information you'll have:
 The user's objective: This is the task you're trying to
 ↳ complete.
 The web page's accessibility tree: This is a simplified
 ↳ representation of the webpage, providing key
 ↳ information.
 The web page's URL: This is the page you're currently
 ↳ navigating.
 The open tabs: These are the tabs you have open.

 The actions you can perform fall into several categories
 ↳ :

 Page Operation Actions:
 'click [id]': This action clicks on an element with a
 ↳ specific id on the webpage.
 'type [id] [content] [press_enter_after=0/1]': Use this
 ↳ to type the content into the field with id. By
 ↳ default, the "Enter" key is pressed after typing
 ↳ unless press_enter_after is set to 0.
 'hover [id]': Hover over an element with id.
 'press [key_comb]': Simulates the pressing of a key
 ↳ combination on the keyboard (e.g., Ctrl+v).
 'scroll [direction=down/up]': Scroll the page up or down
 ↳ .

 Tab Management Actions:
 'new_tab': Open a new, empty browser tab.
 'tab_focus [tab_index]': Switch the browser's focus to a
 ↳ specific tab using its index.
 'close_tab': Close the currently active tab.

 URL Navigation Actions:
 'goto [url]': Navigate to a specific URL.
 'go_back': Navigate to the previously viewed page.
 'go_forward': Navigate to the next page (if a previous '
 ↳ go_back' action was performed).

 Completion Action:
 'stop [answer]': Issue this action when you believe the
 ↳ task is complete. If the objective is to find a
 ↳ text-based answer, provide the answer in the
 ↳ bracket. If you believe the task is impossible
 ↳ to complete, provide the answer as "N/A" in the
 ↳ bracket.

 Now you are trying to evaluate your performance on a
 ↳ past task. You will be given the objective of
 ↳ the task, the history of interaction including
 ↳ the observations you had and the actions you
 ↳ issued, and the status of the task. You will
 ↳ also be given the memory of your previous
 ↳ attempts. Your goal is to think about the
 ↳ strategy and path you took to attempt to
 ↳ complete the task. Try to summarize the reason
 ↳ why you failed to complete the task, and devise
 ↳ a concise, new plan that accounts for your
```

```
↳ mistake and can be helpful when you are solving
↳ the same task. Try to think differently from the
↳ previous attempts. Try to focus on the key
↳ aspect and make the plan concise.
""",
 "examples": [
 (
 ""OBJECTIVE: Compare the time for walking and
 ↳ driving route from AMC Waterfront to Carnegie
 ↳ Mellon University
```

#### OBSERVATION AND ACTION HISTORY:

OBSERVATION 0:

Tab 0 (current): Dashboard / Magento Admin

```
[1] RootWebArea 'Dashboard / Magento Admin' focused:
 ↳ True
[178] link 'Magento Admin Panel'
[201] img 'Magento Admin Panel'
[85] menubar '' orientation: horizontal
[87] link '\ue604 DASHBOARD'
[90] link '\ue60b SALES'
[96] link '\ue608 CATALOG'
[102] link '\ue603 CUSTOMERS'
[108] link '\ue609 MARKETING'
[114] link '\ue602 CONTENT'
[120] link '\ue60a REPORTS'
[138] link '\ue60d STORES'
[144] link '\ue610 SYSTEM'
[150] link '\ue612 FIND PARTNERS & EXTENSIONS'
[821] button 'System Messages: 1'
[902] StaticText 'One or more '
[903] link 'indexers are invalid'
[904] StaticText '. Make sure your '
[905] link 'Magento cron job'
[906] StaticText ' is running.'
[240] heading 'Dashboard'
[242] link '\ue600 admin'
[244] link '\ue607'
[913] textbox '\ue60c' required: False
[48] main ''
[219] StaticText 'Scope:'
[250] button 'All Store Views' hasPopup: menu
[253] link '\ue633 What is this?'
[226] button 'Reload Data'
[917] HeaderAsNonLandmark ''
[919] StaticText 'Advanced Reporting'
[920] StaticText "Gain new insights and take command
 ↳ of your business' performance, using our
 ↳ dynamic product, order, and customer reports
 ↳ tailored to your customer data."
[921] link 'Go to Advanced Reporting \ue644'
[924] StaticText 'Chart is disabled. To enable the
 ↳ chart, click '
[925] link 'here'
[1154] StaticText 'Revenue'
[1054] StaticText '$0.00'
[1155] StaticText 'Tax'
[1156] StaticText 'Shipping'
[1157] StaticText 'Quantity'
[1068] StaticText '0'
[57] tablist '' multiselectable: False orientation:
 ↳ horizontal
[59] tab 'The information in this tab has been
 ↳ changed. This tab contains invalid data. Please
 ↳ resolve this before saving. Loading...
 ↳ Bestsellers' expanded: True selected: True
 ↳ controls: grid_tab_ordered_products_content
[67] link 'The information in this tab has been
 ↳ changed. This tab contains invalid data. Please
 ↳ resolve this before saving. Loading...
 ↳ Bestsellers'
[61] tab 'The information in this tab has been
 ↳ changed. This tab contains invalid data. Please
 ↳ resolve this before saving. Loading... Most
 ↳ Viewed Products' expanded: False selected: False
 ↳ controls: ui-id-1
[69] link 'The information in this tab has been
 ↳ changed. This tab contains invalid data. Please
```

```

→ resolve this before saving. Loading... Most
→ Viewed Products'
[63] tab 'The information in this tab has been
→ changed. This tab contains invalid data. Please
→ resolve this before saving. Loading... New
→ Customers' expanded: False selected: False
→ controls: ui-id-2
[71] link 'The information in this tab has been
→ changed. This tab contains invalid data. Please
→ resolve this before saving. Loading... New
→ Customers'
[65] tab 'The information in this tab has been
→ changed. This tab contains invalid data. Please
→ resolve this before saving. Loading... Customers
→ ' expanded: False selected: False controls: ui-
→ id-3
[73] link 'The information in this tab has been
→ changed. This tab contains invalid data. Please
→ resolve this before saving. Loading... Customers
→ '
[79] tabpanel 'The information in this tab has been
→ changed. This tab contains invalid data. Please
→ resolve this before saving. Loading...
→ Bestsellers'
[1088] table ''
[1158] row ''
[1159] columnheader 'Product' required: False
[1160] columnheader 'Price' required: False
[1161] columnheader 'Quantity' required: False
[1162] row 'http://localhost:7780/admin/catalog/
→ product/edit/id/29/'
[1167] gridcell 'Sprite Stasis Ball 65 cm'
→ required: False
[1168] gridcell '$27.00' required: False
[1169] gridcell '6' required: False
[930] StaticText 'Lifetime Sales'
[933] StaticText '$0.00'
[937] StaticText 'Average Order'
[944] StaticText 'Last Orders'
[945] table ''
[979] row ''
[980] columnheader 'Customer' required: False
[981] columnheader 'Items' required: False
[982] columnheader 'Total' required: False
[983] row 'http://localhost:7780/admin/sales/order
→ /view/order_id/299/'
[988] gridcell 'Sarah Miller' required: False
[989] gridcell '5' required: False
[990] gridcell '$194.40' required: False
[984] row 'http://localhost:7780/admin/sales/order
→ /view/order_id/65/'
[991] gridcell 'Grace Nguyen' required: False
[992] gridcell '4' required: False
[993] gridcell '$190.00' required: False

```

ACTION 0: stop [N/A]

STATUS: FAILED

REFLECTIONS FROM PREVIOUS ATTEMPTS: none""",

```

"I think the task is impossible to complete, thus
→ I issue the stop action. However, the task is
→ not completed successfully, which means I am
→ wrong. I think I should go to the \"REPORT\" tab
→ and do a search there for the best-selling
→ products next time."
),
(

```

```

""OBJECTIVE: List out reviewers, if exist, who
→ mention about good fingerprint resistant

```

OBSERVATION AND ACTION HISTORY:

OBSERVATION 0:

```

URL: http://localhost:7770/3-pack-samsung-galaxy-s6-
→ screen-protector-nearpow-tempered-glass-screen-
→ protector-with-9h-hardness-crystal-clear-easy-
→ bubble-free-installation-scratch-resist.html

```

```

Tab 0 (current): [3 Pack] Samsung Galaxy S6 Screen
→ Protector, Nearpow [Tempered Glass] Screen

```

```

→ Protector with [9H Hardness] [Crystal Clear] [
→ Easy Bubble-Free Installation] [Scratch Resist]
[1] RootWebArea '[3 Pack] Samsung Galaxy S6 Screen
→ Protector, Nearpow [Tempered Glass] Screen
→ Protector with [9H Hardness] [Crystal Clear] [
→ Easy Bubble-Free Installation] [Scratch Resist]'
→ focused: True
[1314] link 'My Account'
[1312] link 'My Wish List'
[1316] link 'Sign Out'
[1319] StaticText 'Welcome, Emma Lopez!'
[1220] link 'Skip to Content'
[1229] link 'store logo'
[1322] img 'one_stop_market_logo'
[1323] link '\ue611 My Cart'
[2246] StaticText 'Search'
[1508] combobox '\ue615 Search' autocomplete: both
→ hasPopup: listbox required: False expanded:
→ False
[2249] link 'Advanced Search'
[1511] button 'Search' disabled: True
[1096] tablist '' multiselectable: False orientation:
→ horizontal
[1098] tabpanel ''
[40] menu '' orientation: vertical
[791] menuitem '\ue622 Beauty & Personal Care'
→ hasPopup: menu
[856] menuitem '\ue622 Sports & Outdoors'
→ hasPopup: menu
[866] menuitem '\ue622 Clothing, Shoes & Jewelry
→ ' hasPopup: menu
[880] menuitem '\ue622 Home & Kitchen' hasPopup:
→ menu
[917] menuitem '\ue622 Office Products' hasPopup
→ : menu
[925] menuitem '\ue622 Tools & Home Improvement'
→ hasPopup: menu
[930] menuitem '\ue622 Health & Household'
→ hasPopup: menu
[936] menuitem '\ue622 Patio, Lawn & Garden'
→ hasPopup: menu
[941] menuitem '\ue622 Electronics' hasPopup:
→ menu
[1002] menuitem '\ue622 Cell Phones &
→ Accessories' hasPopup: menu
[1017] menuitem '\ue622 Video Games' hasPopup:
→ menu
[1030] menuitem '\ue622 Grocery & Gourmet Food'
→ hasPopup: menu
[1253] link 'Home'
[1256] StaticText '[3 Pack] Samsung Galaxy S6 Screen
→ Protector, Nearpow [Tempered Glass] Screen
→ Protector with [9H Hardness] [Crystal Clear] [
→ Easy Bubble-Free Installation] [Scratch Resist]'
[5] main ''
[1257] heading '[3 Pack] Samsung Galaxy S6 Screen
→ Protector, Nearpow [Tempered Glass] Screen
→ Protector with [9H Hardness] [Crystal Clear] [
→ Easy Bubble-Free Installation] [Scratch Resist]'
[11] generic 'Availability'
[13] StaticText 'IN STOCK'
[1331] StaticText 'SKU'
[1467] StaticText 'B01G31IYMO'
[1264] LayoutTable ''
[1469] StaticText 'Rating:'
[1334] generic '78%'
[2221] StaticText '% of'
[2224] StaticText '100'
[1335] link '12\xa0Reviews '
[1336] link 'Add Your Review'
[1338] StaticText '$7.99'
[1279] LayoutTable ''
[1483] StaticText 'Qty'
[1484] spinbutton 'Qty' required: False valuemin:
→ 0 valuemax: 0 valuetext:
[1485] button 'Add to Cart'
[1281] link 'Add to Wish List'
[1282] link 'Add to Compare'
[1287] link 'Skip to the end of the images gallery'

```

```
[1117] button 'Previous'
[1119] generic 'Image'
 [2252] img 'Image'
[1118] button 'Next'
```

ACTION 0:

```
click [1335] where [1335] is [1335] link '12\xa0Reviews'
 ↳ '
```

OBSERVATION 1:

```
URL: http://localhost:7770/3-pack-samsung-galaxy-s6-
 ↳ screen-protector-nearpow-tempered-glass-screen-
 ↳ protector-with-9h-hardness-crystal-clear-easy-
 ↳ bubble-free-installation-scratch-resist.html
Tab 0 (current): [3 Pack] Samsung Galaxy S6 Screen
 ↳ Protector, Nearpow [Tempered Glass] Screen
 ↳ Protector with [9H Hardness] [Crystal Clear] [
 ↳ Easy Bubble-Free Installation] [Scratch Resist]
[1] RootWebArea '[3 Pack] Samsung Galaxy S6 Screen
 ↳ Protector, Nearpow [Tempered Glass] Screen
 ↳ Protector with [9H Hardness] [Crystal Clear] [
 ↳ Easy Bubble-Free Installation] [Scratch Resist]'
 ↳ focused: True
[5] main ''
 [1349] StaticText 'Skip to the beginning of the
 ↳ images gallery'
 [1106] tablist '' multiselectable: False orientation
 ↳ : horizontal
 [1107] tab 'Details' expanded: False selected:
 ↳ False controls: description
 [1350] link 'Details'
 [1110] tab 'Reviews (12)' expanded: True selected:
 ↳ True controls: reviews
 [1352] link 'Reviews (12)'
 [2365] tabpanel 'Reviews (12)'
 [2460] StaticText 'Customer Reviews'
 [2555] StaticText "Best screen protectors I've
 ↳ used!"
 [2519] LayoutTable ''
 [2556] LayoutTableRow ''
 [2699] LayoutTableCell 'Rating'
 [2700] generic '100%'
 [2559] StaticText 'It is super clear and
 ↳ fingerprint resistant. It was kind of hard
 ↳ trying to get it on, and I did get some hairs on
 ↳ the sticky side, but all in all it was great!
 ↳ Bubbles went away around the small hairs so you
 ↳ can barely tell they are there. They also give
 ↳ you tons of extra tools to help you clean the
 ↳ screen and get dust particles off of the screen
 ↳ before you put it on. I think it was just me
 ↳ being clumsy with all of the dust particles
 ↳ getting inside the screen.'
 [2562] StaticText 'Review by '
 [2564] StaticText 'Rachel'
 [2567] StaticText 'Posted on '
 [2568] time ''
 [2701] StaticText '4/18/23'
 [2569] StaticText 'Good screen protector for the
 ↳ money and good customer service'
 [2522] LayoutTable ''
 [2570] LayoutTableRow ''
 [2702] LayoutTableCell 'Rating'
 [2703] generic '80%'
 [2573] StaticText 'This is the second time I
 ↳ have used this product. It is a little tricky to
 ↳ apply. I had it on my phone for about 10 months
 ↳ and had dropped my phone a few times without
 ↳ incident. The last drop shattered the protector
 ↳ but thankfully did what it was supposed to do
 ↳ and protected my phone screen. The second one in
 ↳ the package had a small chip in it, which
 ↳ caused it to have a hairline crack all the way
 ↳ through. I emailed the company and they were
 ↳ very quick to respond and sent a new one free of
 ↳ charge. I am very satisfied with the product
 ↳ and only give it a four star rating because it
 ↳ is sometimes very difficult to get out the
 ↳ bubbles. I have 2 very tiny specks that would
```

```
↳ just not come out.'
 [2576] StaticText 'Review by '
 [2578] StaticText 'chris'
 [2581] StaticText 'Posted on '
 [2582] time ''
 [2704] StaticText '4/18/23'
 [2583] StaticText 'Bubbles still there after a
 ↳ few days'
 [2525] LayoutTable ''
 [2584] LayoutTableRow ''
 [2705] LayoutTableCell 'Rating'
 [2706] generic '80%'
 [2587] StaticText "OK, so my first impression
 ↳ was, wow it worked with only 1 small bubble. I
 ↳ was like OK, it's normal to have a few small
 ↳ bubbles. The description says that the small
 ↳ bubbles will disappear after a couple days. Well
 ↳ it's been over a week and the one small tiny
 ↳ bubble is still there. It never went away. Ugh I
 ↳ need to add this to my review. The glue does
 ↳ not last forever. It started to come off about a
 ↳ month after I put it on. The bad thing when it
 ↳ does start to come off, it's easy to take off
 ↳ the screen protectant."
```

ACTION 1:

```
stop [Rachel]
```

STATUS: FAILED

```
REFLECTIONS FROM PREVIOUS ATTEMPTS: none"",
 "I find the review from Rachel, which is the
 ↳ answer to the objective. I issue the stop action
 ↳ with the answer. However, the task is not
 ↳ completed successfully. This might be because I
 ↳ missed other reviews that also mention about
 ↳ good fingerprint resistant. I think I should
 ↳ read all the reviews next time."
),
],
"template": ""OBJECTIVE: {objective}
```

OBSERVATION AND ACTION HISTORY:

```
{trajectory}
STATUS: {status}
```

```
REFLECTIONS FROM PREVIOUS ATTEMPTS: {memory}""",
"meta_data": {
 "observation": "accessibility_tree",
 "action_type": "id_accessibility_tree",
 "keywords": ["objective", "trajectory", "status", "
 ↳ memory"],
 "prompt_constructor": "
 ↳ ReflectionGenerationPromptConstructor",
 "answer_phrase": "",
 "action_splitter": "```"
},
}
```